Building community-specific standards and vocabularies: prospects and challenges for linking to the broader community - The SINP Case

Remy Jomier[‡], Paula F Zermoglio[§], John Wieczorek[|]

‡ Natural History Museum, Paris, France
§ Instituto de Ecología, Genética y Evolución de Buenos Aires (IEGEBA-CONICET), University of Buenos Aires, Buenos Aires, Argentina
| Museum of Vertebrate Zoology, University of California, Berkeley, United States of America

Abstract

Biodiversity data may come from myriad sources. From data capture in the field through digitization processes, each source may choose distinctive ways to capture data. When it comes to sharing data more broadly at national or regional levels, it is imperative that data is presented in ways that encourage understanding both by humans and machines, allowing aggregation and serving the data back to the community. This implies two levels of agreement, one at a structural level, where data is organized under certain terms or fields, and another related to the actual values contained in such fields. Since its ratification in 2009, the Darwin Core standard Wieczorek et al. (2012) has been increasingly used across the community to respond to the first need, providing a relatively simple means to organize shared data. Nonetheless, despite its broad acceptance, efforts to develop different standards to answer the same problems are not uncommon among some stakeholders, and may introduce yet another issue: reconciling the data shared under different standards. The second level of agreement, at the value level, constitutes a much more complex issue, partly given the nature of biodiversity data and partly due to social constraints. Many potential, partial solutions involving the development of dictionaries and

controlled vocabularies are found scattered across the community. As the lack of homogeneity renders data less discoverable (Zermoglio et al. 2016) and therefore less usable for research and decision making, there exists a growing need for unifying such efforts.

As part of the Biodiversity Information System on Nature and Landscapes (SINP), the French National Museum of Natural History was appointed to develop biodiversity data exchange standards, with the goal of sharing French marine and terrestrial data at the national level, meeting national and European requirements (e.g., the European INSPIRE Directive European Commission 2017). The French data providers include a broad range of people with diverse backgrounds. While some stakeholders can provide data under very specific constraints and formats, others lack the capabilities or resources to do so. The variability in the data provided therefore extends through both the structure and the value levels. In order to integrate the data in a coherent national system, a dedicated working group was assembled, mobilizing a range of biodiversity stakeholders and experts. Existing standards were compared, existing vocabularies gathered and compiled for review by experts, and then presented to the working group. As a result, a set of terms and associated controlled vocabularies was established. Finally, the set was released to the public to test and amended as needed.

The results of the French initiative proved useful to compile and share data at the national level, bringing together data providers that otherwise would have been excluded. However, at a global scale, it faces some challenges that still need to be fully addressed. For instance, the standards created do not have an exact correspondence with Darwin Core, and so a complex mapping is required in order to integrate the data with that of the rest of the community. A serious mapping effort is being carried out as the national standards progress and has already rendered good results (Jomier and Pamerlon 2016).

Regardless of the problems that remain to be solved, some lessons can be learnt from this effort. Getting actively involved in the broader, global community can help identify available tools, resources and expertise, and avoid repeated efforts that can be costly and time-consuming. Furthermore, re-using elements that already have been proven to work, prevents the need for reconciliations and makes data integration easier. With the ultimate goal of making biodiversity data readily available, these lessons should be kept in mind for future initiatives.

Keywords

biodiversity data standards, controlled vocabularies, SINP, community engagement

Presenting author

Remy Jomier, Paula Zermoglio

Presented at

TDWG Conference 2017

References

- European Commission (2017) About INSPIRE. http://inspire.ec.europa.eu/about-inspire/563. Accessed on: 2017-7-20.
- Jomier R, Pamerlon S (2016) French national standards for biodiversity: interoperability on an international level. Biodiversity Information Standards (TDWG) 2016 Annual Conference, Costa Rica, 2016. URL: https://mbgserv18.mobot.org/ocs/index.php/tdwg/tdwg2016/paper/view/1039
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: an evolving community-developed biodiversity data standard. PLoS ONE 7 (1): 29715. https://doi.org/10.1371/journal.pone.0029715
- Zermoglio PF, Bloom D, Wieczorek JR, Guralnick RP, LaFrance R, Russell L (2016) Darwin Core Documentation: More (would be) Better. Biodiversity Information Standards (TDWG) 2016 Annual Conference, Costa Rica, 2016. URL: https://mbgserv18.mobot.org/ocs/index.php/tdwg/tdwg2016/paper/view/1076